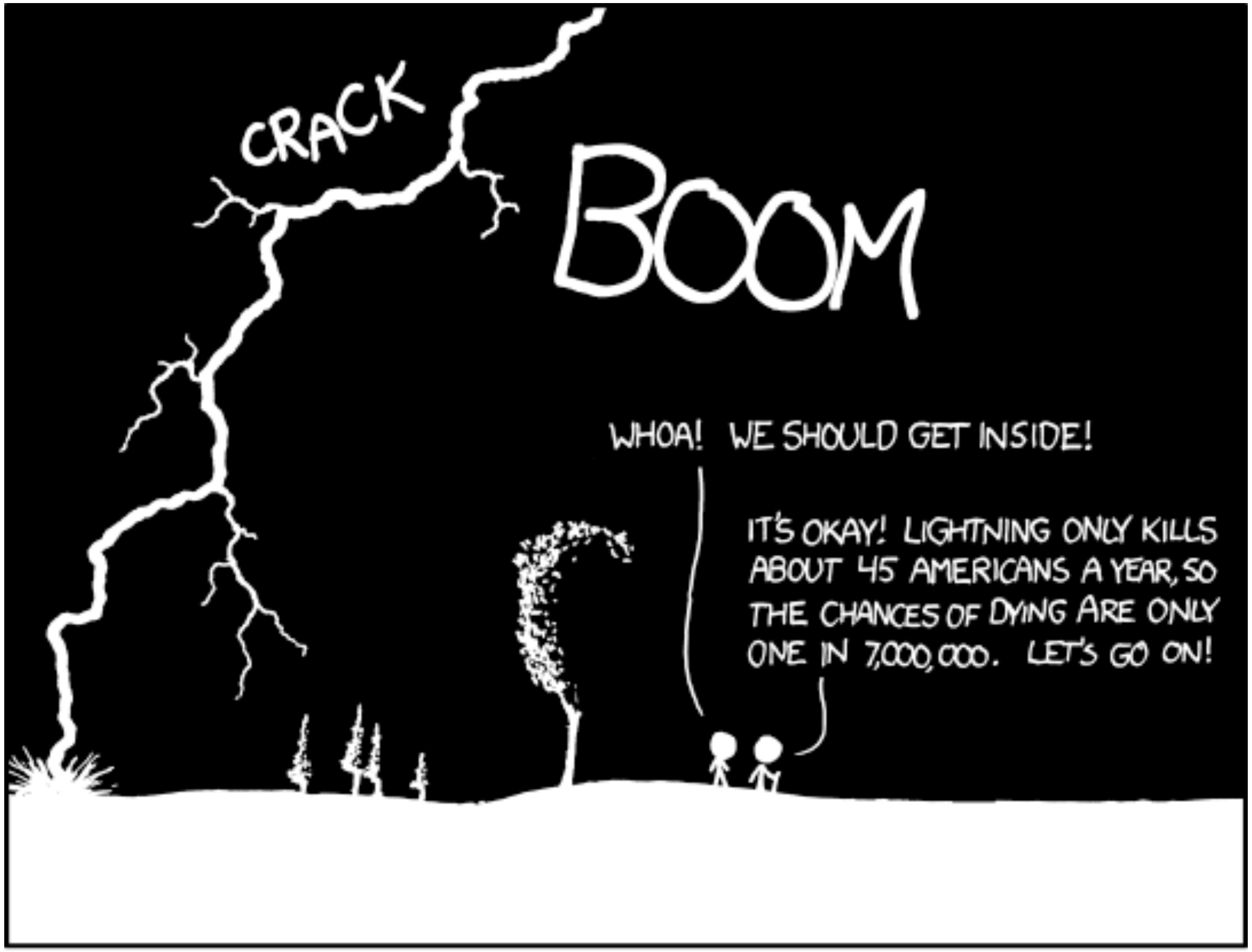
Parameter estimation

Conditional risk



THE ANNUAL DEATH RATE AMONG PEOPLE WHO KNOW THAT STATISTIC IS ONE IN SIX.

Formalizing the problem

- Specify random variables we care about, e.g., Commute Time
- We might then pick a particular distribution over these random variables
  - Say we think our variable is Gaussian
- Now want a way to use data to inform models
  - Let data tell us the parameters for that Gaussian
- **Note**: I do not expect you to be an expert in all the PMFs and PDFs discussed, nor memorize their formulas

Parameter estimation

- Assume that we are given some model class, M,
  - e.g., Gaussian with parameters mu and sigma
  - selection of model from the class corresponds to selecting mu, sigma
- Now want to select "best" model; how do we define best?
  - Generally assume data comes from that model class; might want to find model that best explains the data (or most likely given the data)
  - Might want most likely model, with preference for "important" samples
  - Might want most likely model, that also matches expert prior info
  - Might want most likely model, that is the simplest (least parameters)
- These additional requirements are usually in place to enable better generalization to unseen data

Some notation

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

$$\prod_{i=1}^{n} x_i = x_1 x_2 \ldots x_n$$

$$c : \mathbb{R}^d \to \mathbb{R}$$

$$\mathbf{w}^* = \arg \max_{\mathbf{w} \in \mathbb{R}^d} c(\mathbf{w})$$

$$c(\mathbf{w}^*) = \max_{\mathbf{w} \in \mathbb{R}^d} c(\mathbf{w})$$

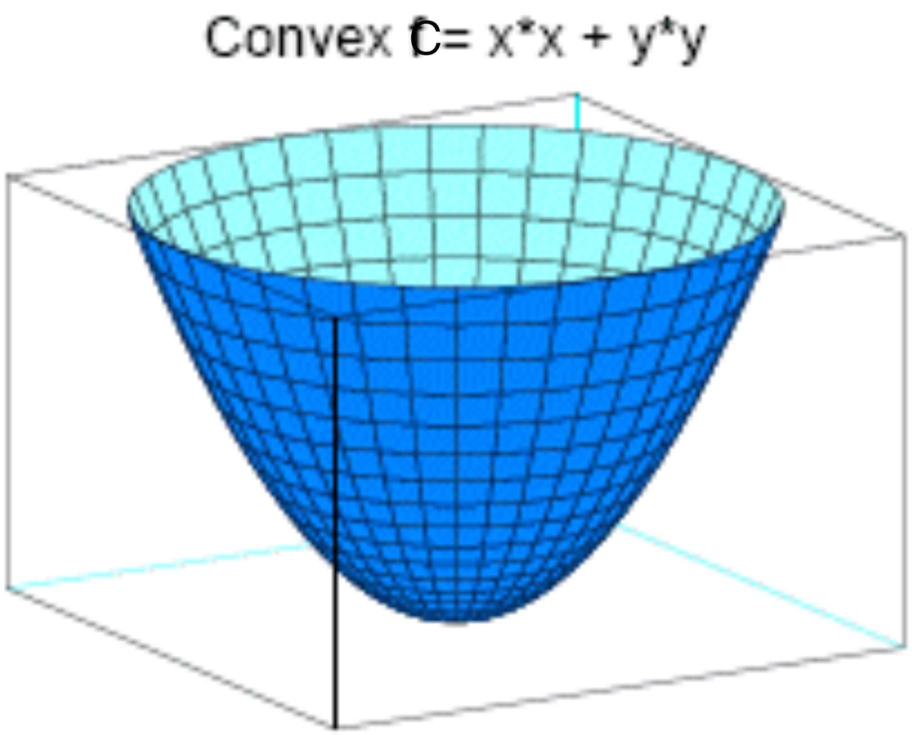
$\mathcal{F}$ is a set of models

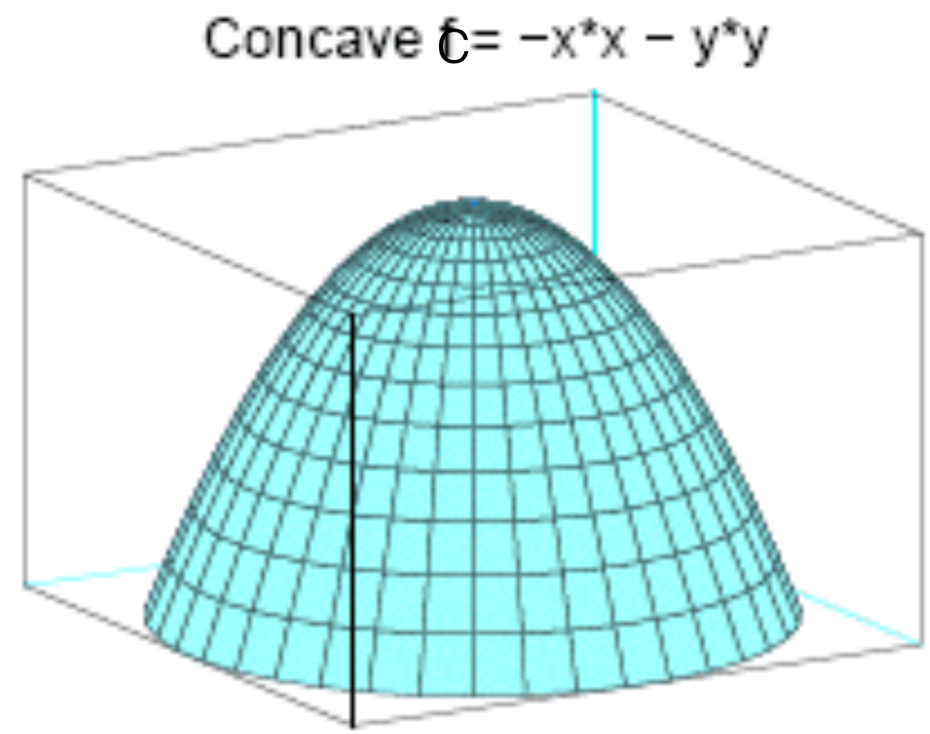
e.g., $\mathcal{F} = \{\mathcal{N}(\mu, \sigma) \mid (\mu, \sigma) \in \mathbb{R}^2, \sigma > 0\}$

e.g., $\mathcal{F} = \{\mathbf{w} \in \mathbb{R}^d \mid f(\mathbf{x}) = \mathbf{x}^\top \mathbf{w}\}$

Definition of optimization

- We select some (error) function c we care about





- Maximizing c means we are finding largest point

- Minimizing c means we are finding smallest point

Maximum a posteriori (MAP) estimation

$$f_{\mathrm{MAP}} = \arg\max_{f \in \mathcal{F}} p(f|\mathcal{D})$$

- Want the f that is most likely, given the data
- p(f | D) is the **posterior distribution** of the model given data
  - e.g., F could be the space of Gaussian distributions, the model is f and f(x) returns probability/density of a point x
  - e.g., we could assume x is Gaussian distributed with variance = 1, and so F could be the reals, and the model f is the mean

Question: What is the function we are optimizing and what are the parameters we are learning?

MAP

$$f_{\text{MAP}} = \arg\max_{f \in \mathcal{F}} p(f|\mathcal{D})$$

- p(f I D) is the **posterior distribution** of the model given data
  - e.g., we could assume x is Gaussian distributed with variance = 1, and so F could be the reals, and the model f is the mean

Question: What is the function we are optimizing and what are the parameters we are learning?

$$c(f) = p(\text{mean is } f|\mathcal{D})$$

$$\max_{f \in \mathbb{R}} c(f)$$

Maximum a posteriori (MAP)

$$f_{\text{MAP}} = \arg\max_{f \in \mathcal{F}} p(f|\mathcal{D})$$

- p(f I D) is the **posterior distribution** of the model given data

- In discrete spaces: p(f I D) is the PMF
  - the MAP estimate is exactly the most probable model
  - e.g., bias of coin is 0.1, 0.5, or 0.7, p(f = 0.1 I D), …

- In continuous spaces: p(f I D) is the PDF
  - the MAP estimate is the model with the largest value of the posterior density function
  - e.g., bias of a coin is in [0, 1]

- But what is p(f I D)? Do we pick it? If so, how?

MAP calculation

- Start by applying Bayes rule

$$p(f|\mathcal{D}) = \frac{p(\mathcal{D}|f)p(f)}{p(\mathcal{D})}$$

- p(D I f) is the **likelihood** of the data, under the model
- p(f) is the **prior** of the model
- p(D) is the marginal distribution of the data
  - we will often be able to ignore this term

Why is this conversion important?

- Do not always have a known form for p(f | D)
- We usually have chosen (known) forms for p(D | f) and p(f)
- Let theta = parameters of model (distribution); interchangeable use p(D | f) = p(D | theta) and p(f) = p(theta)
- **Example**: Let D = {x1} (one sample). Then one common choice is a Gaussian over x1: p(D | f) = p(x1 | mu, sigma)
  - p(f | D) is not obvious, since specified our model class for P(D | f)
  - What is p(f) in this case? We may put some prior "preferences" on mu and sigma, e.g., normal distribution around mu, specifying that really large magnitude values in mu are unlikely
  - Specifying and using p(f) is related to regularization and Bayesian parameter estimation, which will will discuss more later

Why is this conversion important?

- Do not always have a known form for p(f | D)
- We usually have chosen (known) forms for p(D | f) and p(f)
- Let theta = parameters of model (distribution); interchangeable use p(D | f) = p(D | theta) and p(f) = p(theta)
- **Example**: Let D = {x1} (one sample). Then one common choice is a Gaussian over x1: p(D | f) = p(x1 | mu, sigma)
  - p(f | D) is not obvious, since specified our model class for P(D | f)
  - What is p(f) in this case? We may put some prior "preferences" on mu and sigma, e.g., normal distribution around mu, specifying that really large magnitude values in mu are unlikely
  - Specifying and using p(f) is related to regularization and Bayesian parameter estimation, which will will discuss more later

How do we compute p(D)?

- If we have p(D, f), can we obtain p(D)?
  - Marginalization

$$p_{X_i}(x_i) = \sum_{x_1} \cdots \sum_{x_{i-1}} \sum_{x_{i+1}} \cdots \sum_{x_k} p_{\mathbf{X}}(x_1, \ldots, x_k)$$

$$p_{X_i}(x_i) = \int_{x_1} \cdots \int_{x_{i-1}} \int_{x_{i+1}} \cdots \int_{x_k} p_{\mathbf{X}}(\boldsymbol{x})\, dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_k$$

- If we have p(D|f) and p(f), do we have p(D, f)?

Data marginal

- Using the formula of total probability

$$p(\mathcal{D}) = \begin{cases} \sum_{f \in \mathcal{F}} p(\mathcal{D}|f)p(f) & f : \text{discrete} \\ \\ \int_{\mathcal{F}} p(\mathcal{D}|f)p(f)df & f : \text{continuous} \end{cases}$$

- Fully expressible in terms of likelihood and prior

Optimization to get model

$$\begin{aligned}\theta_{\text{MAP}} &= \arg\max_{\theta} \frac{P(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\ &= \frac{1}{p(\mathcal{D})} \arg\max_{\theta} P(\mathcal{D}|\theta)p(\theta) \\ &= \arg\max_{\theta} P(\mathcal{D}|\theta)p(\theta)\end{aligned}$$

Will often write:

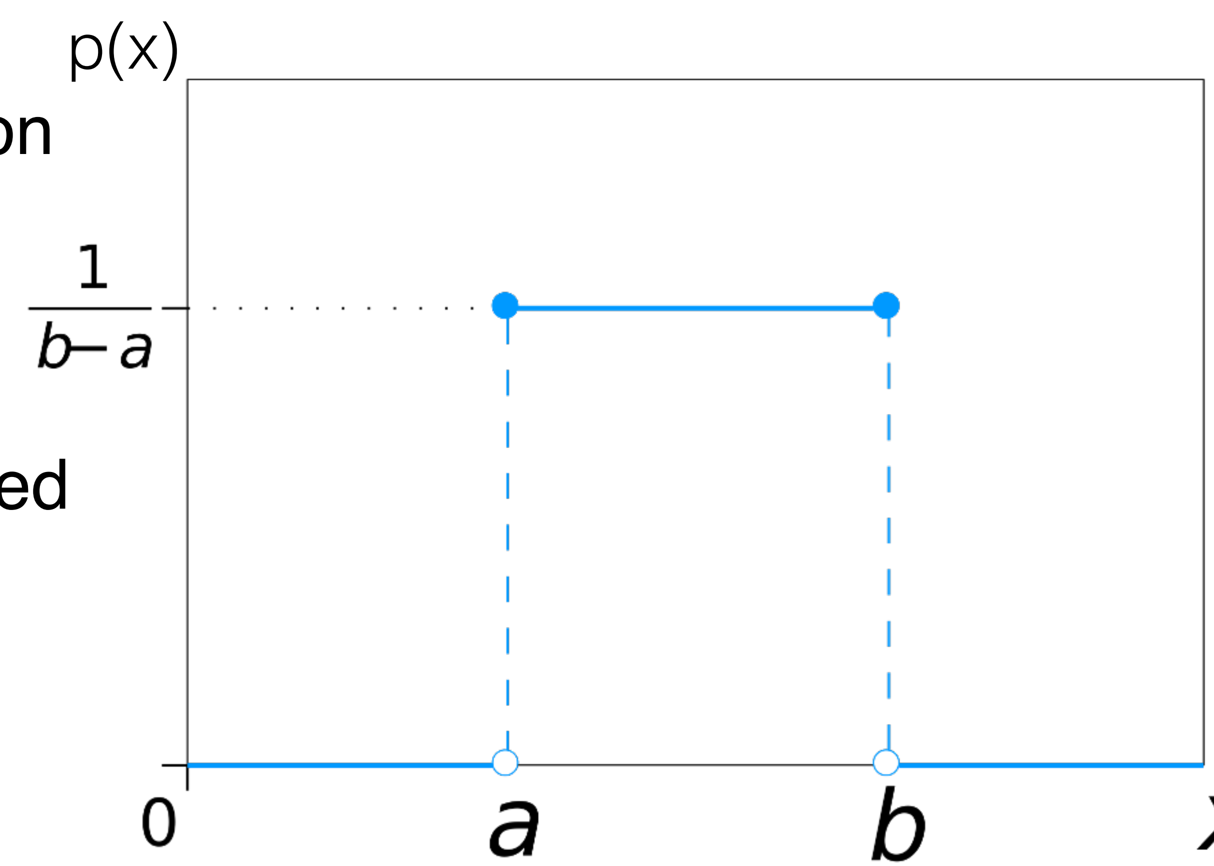
$$\begin{aligned}p(\theta|\mathcal{D}) &= \frac{P(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\ &\propto P(\mathcal{D}|\theta)p(\theta)\end{aligned}$$

Maximum likelihood

$$\theta_{\mathrm{ML}} = \arg\max_{\theta} P(\mathcal{D}|\theta)$$

- In some situations, may not have a reason to prefer one model over another (i.e., no prior knowledge or preferences)

- Can loosely think of maximum likelihood as instance of MAP, with uniform prior p(theta) = u for some constant u
  - If domain is infinite (example, the set of reals), the uniform distribution is not defined!
  - but the interpretation is still similar
  - in practice, typically have a bounded space in mind for the model class

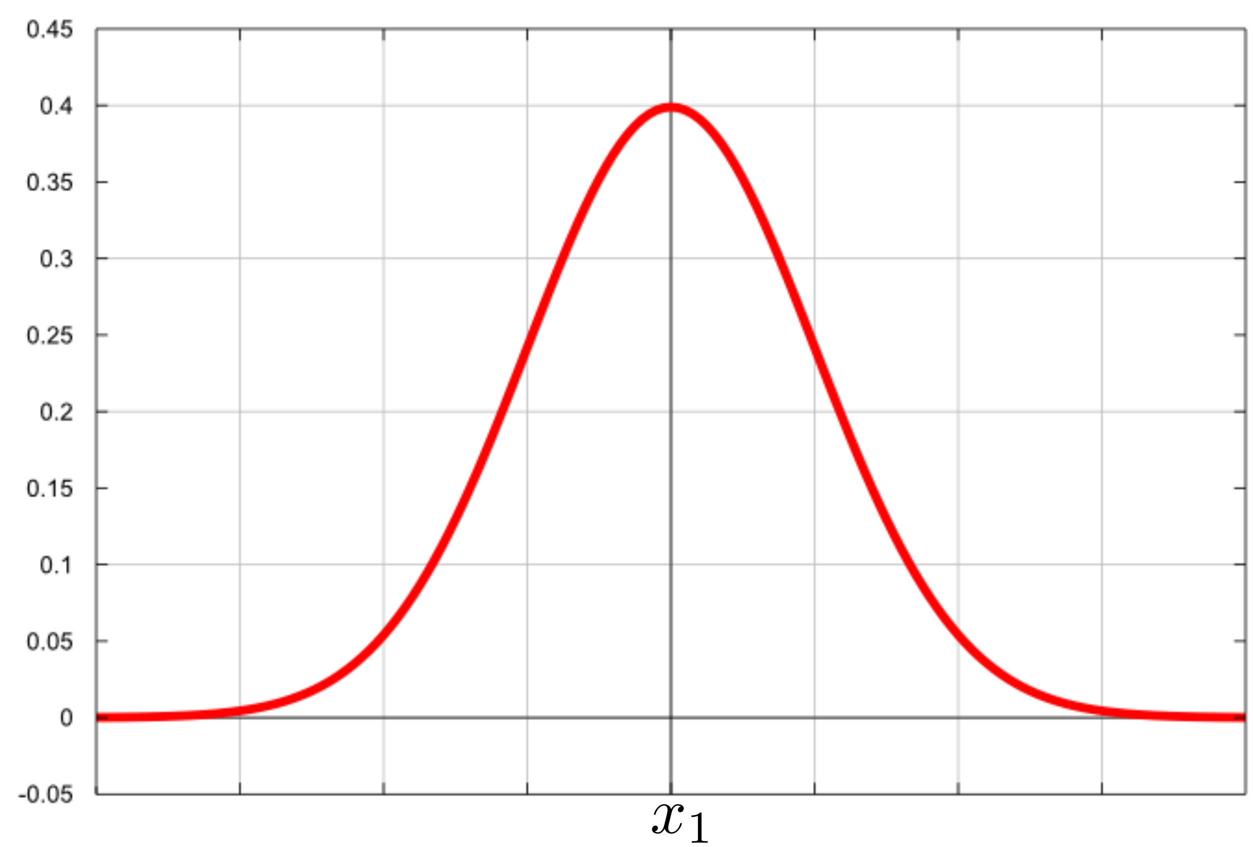


ML example

e.g., $\mathcal{F} = \mathbb{R}, \theta$ is the mean of a Gaussian, fixed $\sigma = 1$

$$
\begin{aligned}
c(\theta) &= p(\mathcal{D}|\theta) \\
&= \mathcal{N}(x_1|\mu = \theta, \sigma^2 = 1) \\
&= \frac{1}{2\pi} \exp\left(-\frac{1}{2}(x_1 - \theta)^2\right)
\end{aligned}
$$

$$c(\theta) = p(\mathcal{D}|\theta)$$



θ

Maximizing the log-likelihood

- We want to maximize the **likelihood**, but often instead maximize the **log-likelihood**

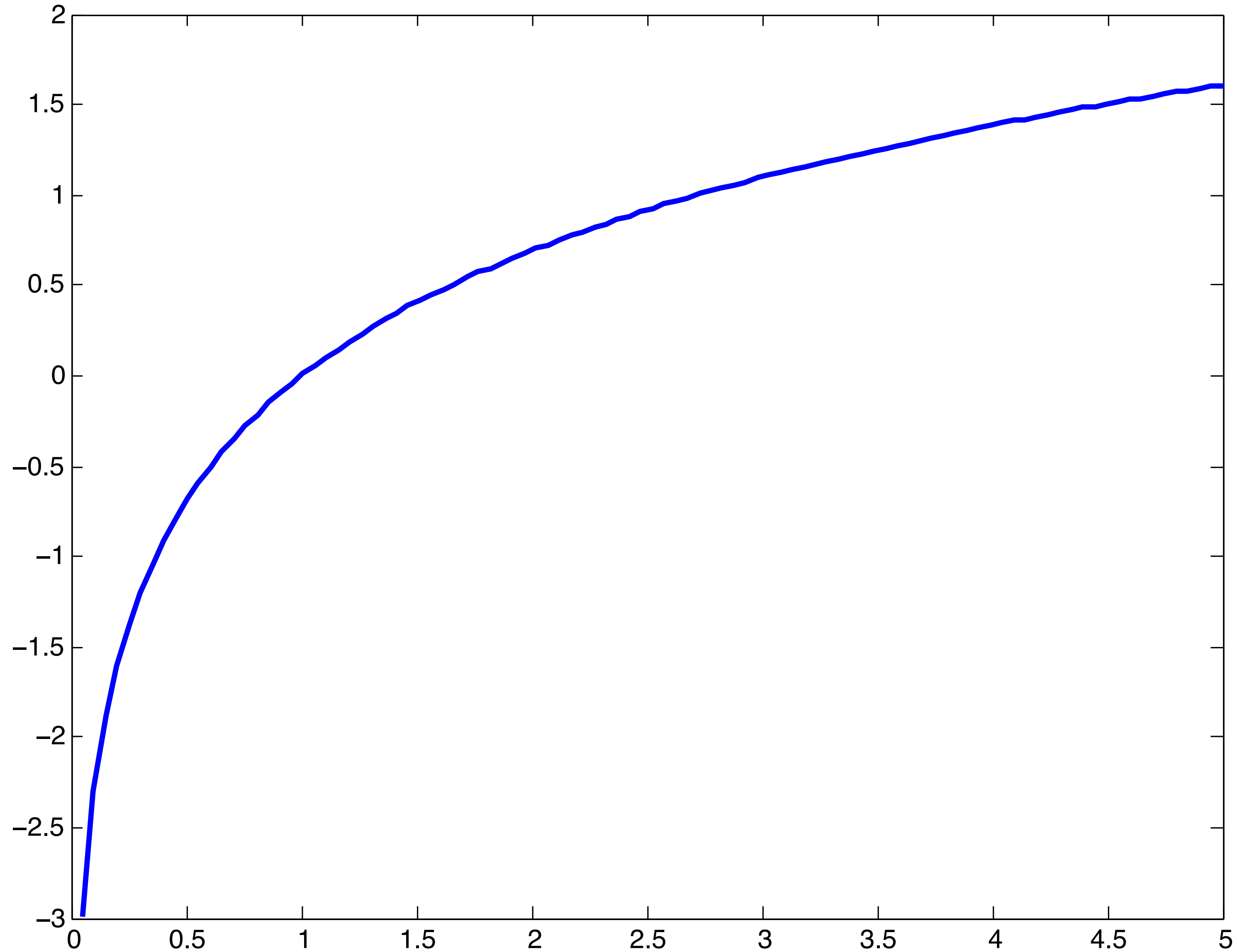
$$\arg\max_{\theta \in \mathcal{F}} p(\mathcal{D}|\theta) = \arg\max_{\theta \in \mathcal{F}} \log p(\mathcal{D}|\theta)$$

- Why? Or maybe first, is this equivalent?
  - The Why is that it makes the optimization much simpler, when we have more than one sample

Why can we shift by log?

$$c(\theta_1) > c(\theta_2) \iff \log c(\theta_1) > \log c(\theta_2)$$

Monotone increasing



Likelihood values always > 0

Maximizing the log-likelihood

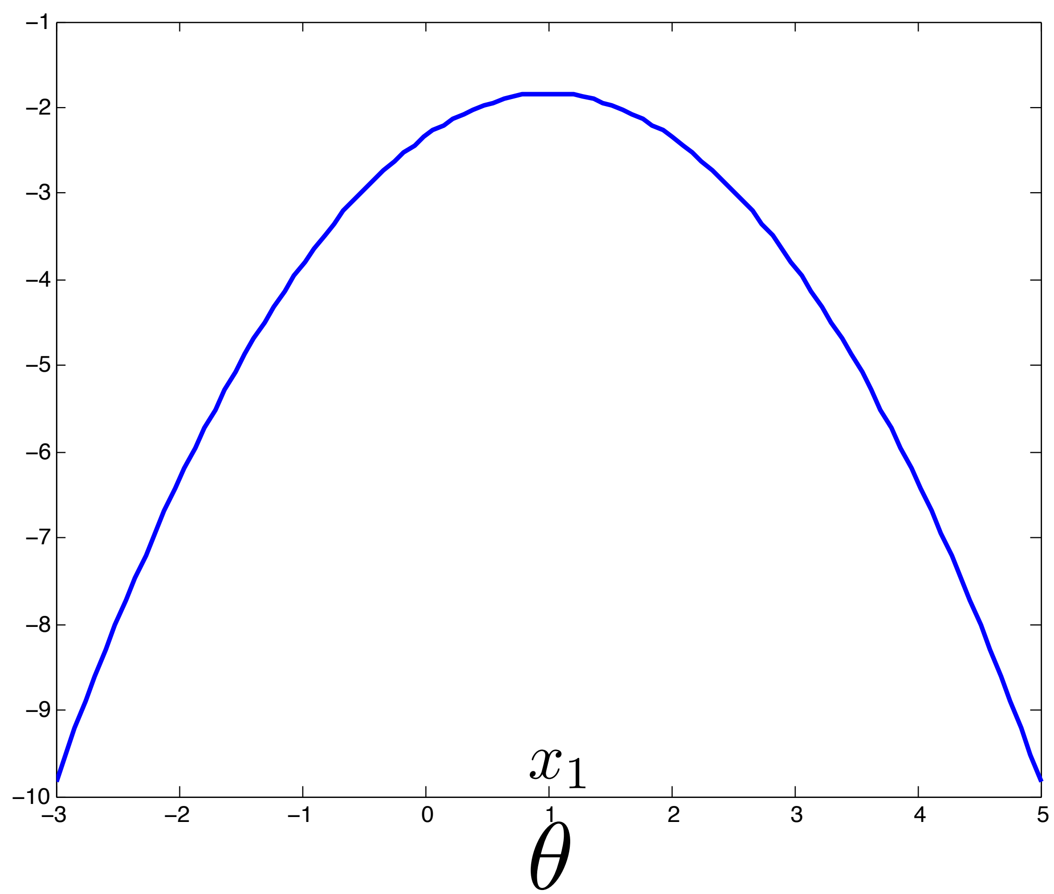
$$\text{e.g., } \mathcal{F} = \mathbb{R}, \theta \text{ is the mean of a Gaussian, fixed } \sigma = 1$$

$$\log(ab) = \log a + \log b$$

$$\log(a^c) = c \log a$$

$$\begin{aligned} c(\theta) &= \log p(\mathcal{D}|\theta) \\ &= \log\left(\frac{1}{2\pi}\exp\left(-\frac{1}{2}(x_1-\theta)^2\right)\right) \\ &= -\log(2\pi) - \frac{1}{2}(x_1-\theta)^2 \end{aligned}$$

$$c(\theta) = \log p(\mathcal{D}|\theta)$$



This conversion is even more important when we have more than one sample

- **Example**: Let D = {x1,x2} (two samples).

- If x1 and x2 are independent samples from same distribution (same model), then P(x1, x2 I theta) = P(x1 I f) P(x2 I theta)

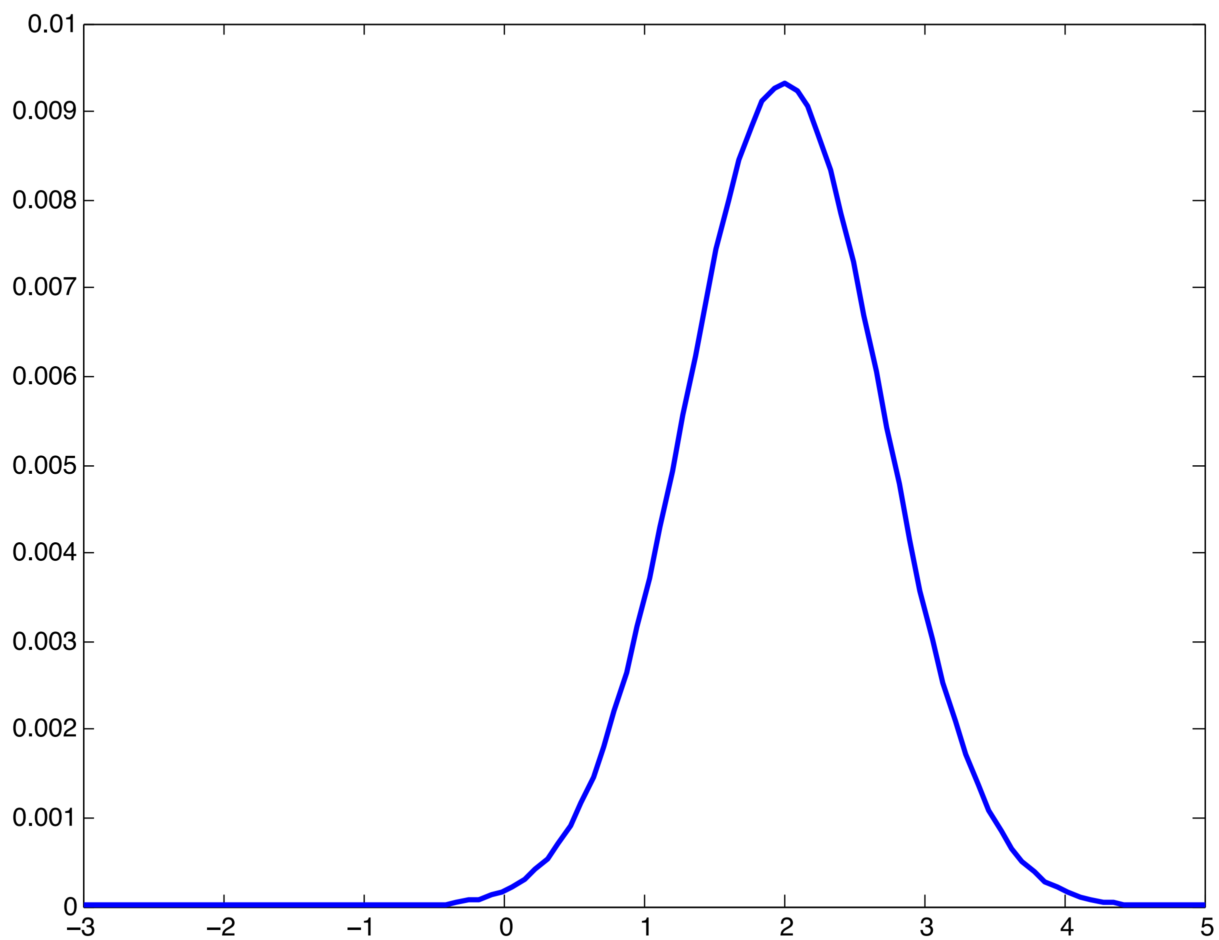
$$p(x_1|\theta)p(x_2|\theta) = \frac{1}{2\pi}\exp\left(-\frac{1}{2}(x_1-\theta)^2\right) \times \frac{1}{2\pi}\exp\left(-\frac{1}{2}(x_2-\theta)^2\right)$$

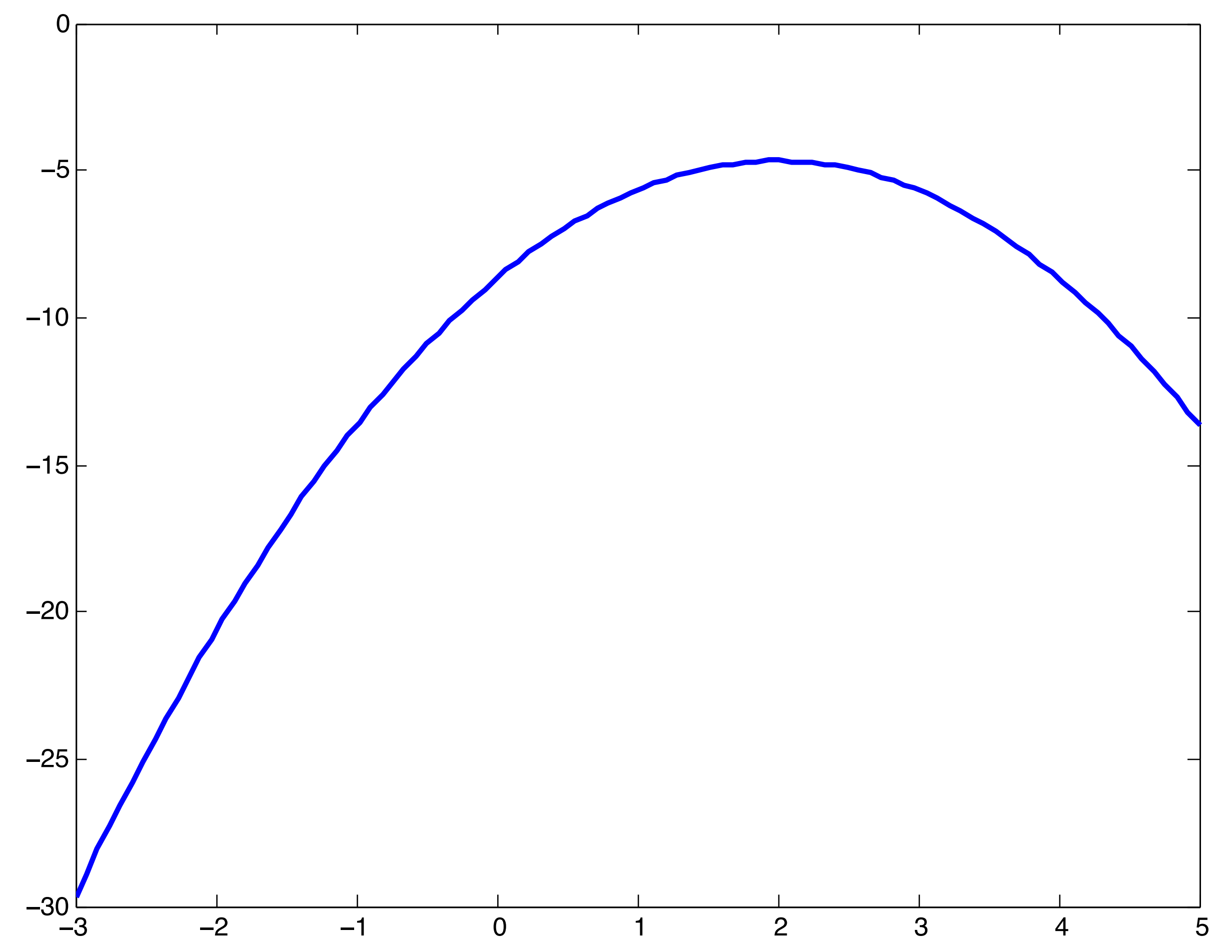
$$\begin{aligned}\log\left(p(x_1|\theta)p(x_2|\theta)\right) &= \log p(x_1|\theta) + \log p(x_2|\theta) \\ &= -2\log(2\pi) - \frac{1}{2}(x_1-\theta)^2 - \frac{1}{2}(x_2-\theta)^2\end{aligned}$$

This conversion is even more important when we have more than one sample

- **Example**: Let D = {x1,x2} (two samples).

- If x1 and x2 are independent samples from same distribution (same model), p(x1, x2 | theta) = p(x1 | theta) p(x2 | theta)





With n samples

- For many iid samples x1, …, xn, we could choose (e.g.,) a Gaussian distribution for P(xi | theta), with theta = {mu,sigma}
    - iid = independent and identically distributed
- P(x1, …, xn | theta) = P(x1 | theta) … P(xn | theta)

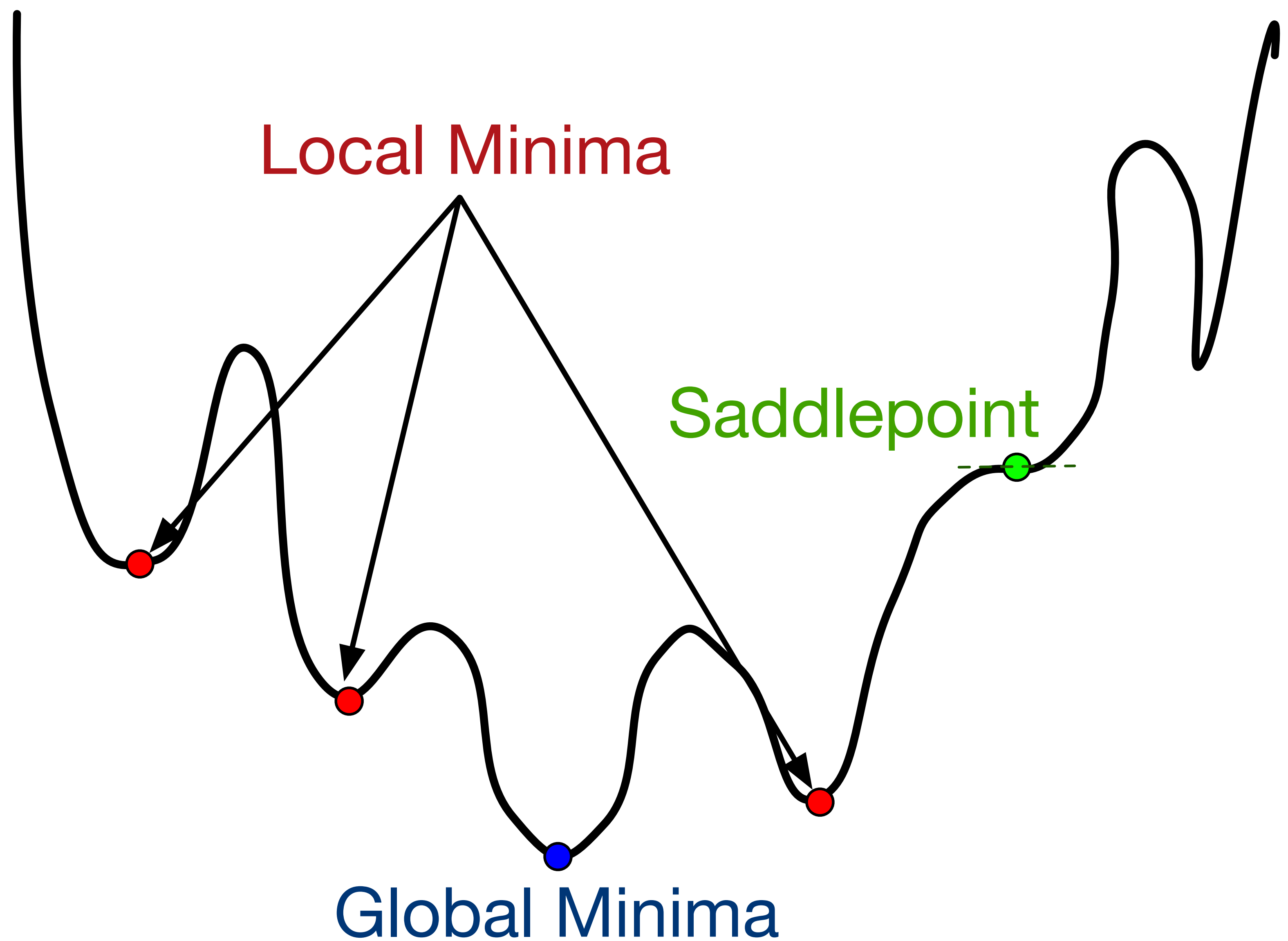
How do we solve this maximization problem?

- Naive strategy:
- 1. Guess 100 solutions theta
- 2. Pick the one with the largest value
- Can we do something better?

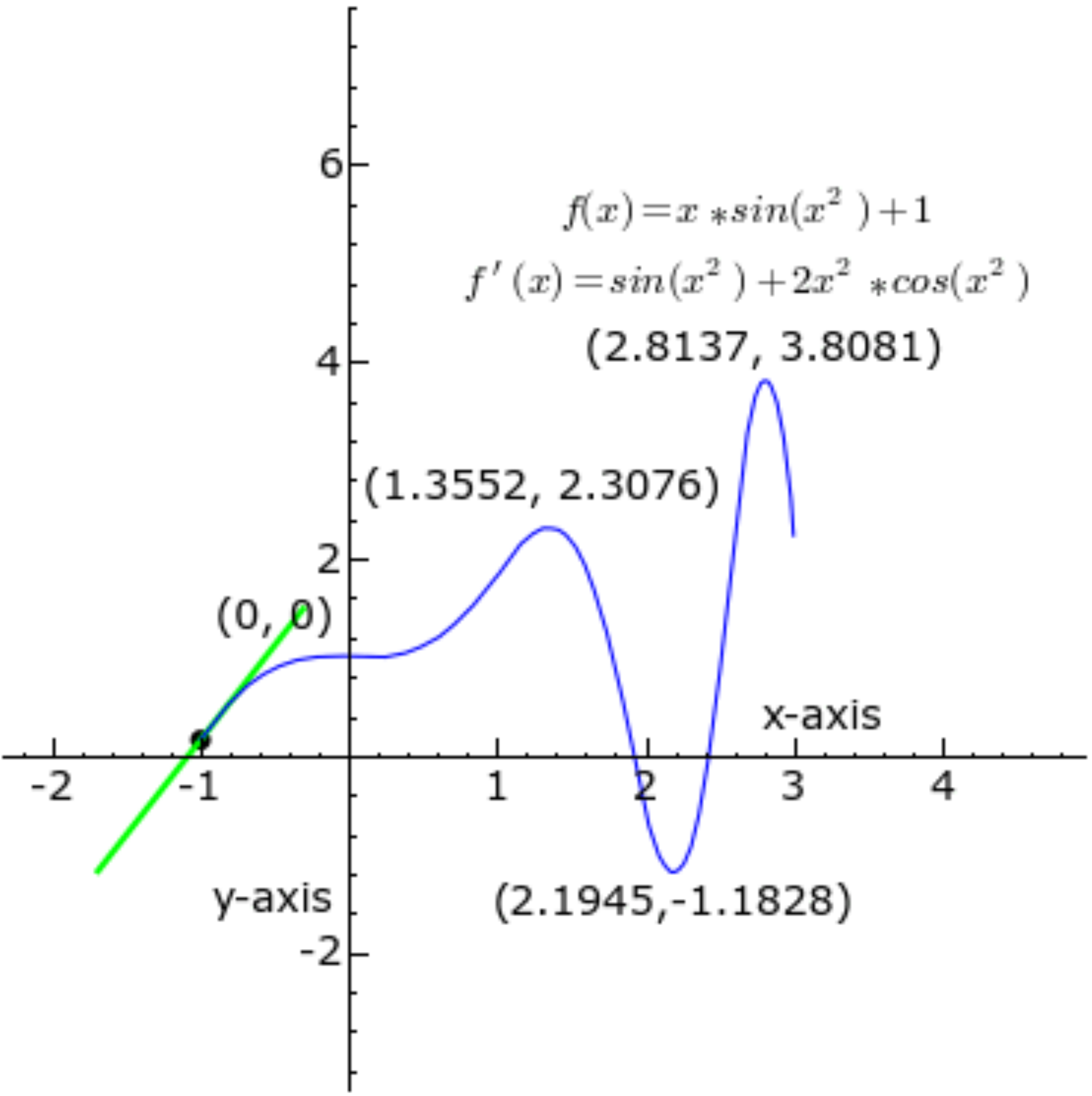
Crash course in optimization

- Goal: find maximal (or minimal) points of functions
- Generally assume functions are smooth, use gradient descent
- Derivative: direction of ascent from a scalar point $\frac{d}{dx}c(x)$
- Gradient: direction of ascent from a vector point $\nabla c(\mathbf{x})$

Function surface



Single-variate calculus



GIF from Wikipedia: Tangent

For a function f defined on a scalar x, the derivative is

$$\frac{df}{dx}(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}$$

At any point, x, $\frac{df}{dx}(x)$ gives the slope of the tangent to the function at f(x)

*Note: f is c in this slide

Why don't constants matter?

$$\max_x c(x)$$

$$\frac{d}{dx} c(x) = 0$$

$$\max_x u c(x), \quad u > 0$$

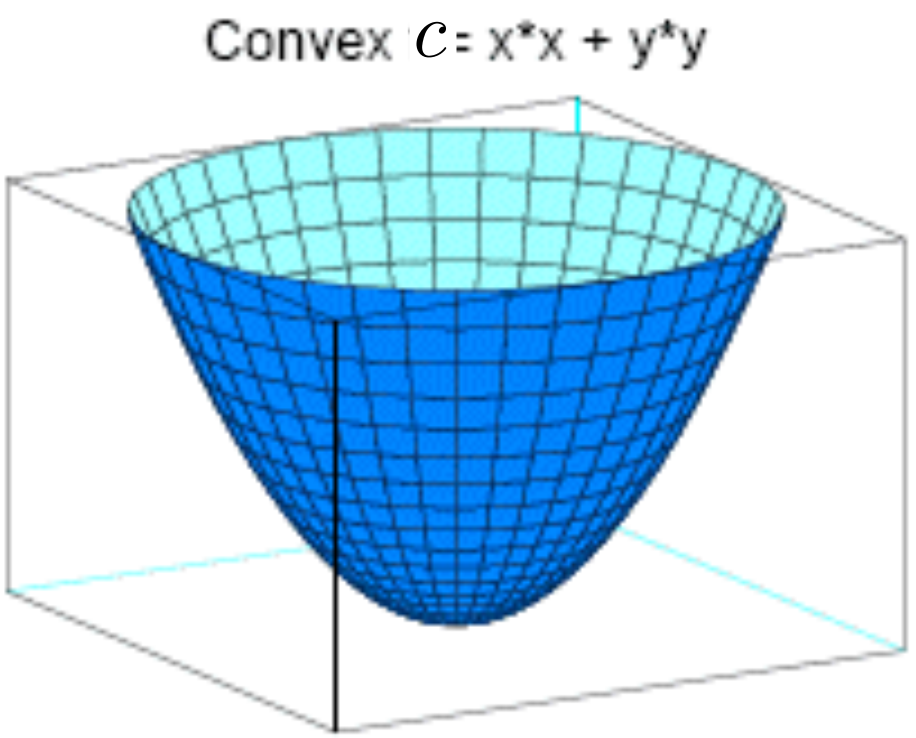
$$\frac{d}{dx} u c(x) = u \frac{d}{dx} c(x) = 0$$

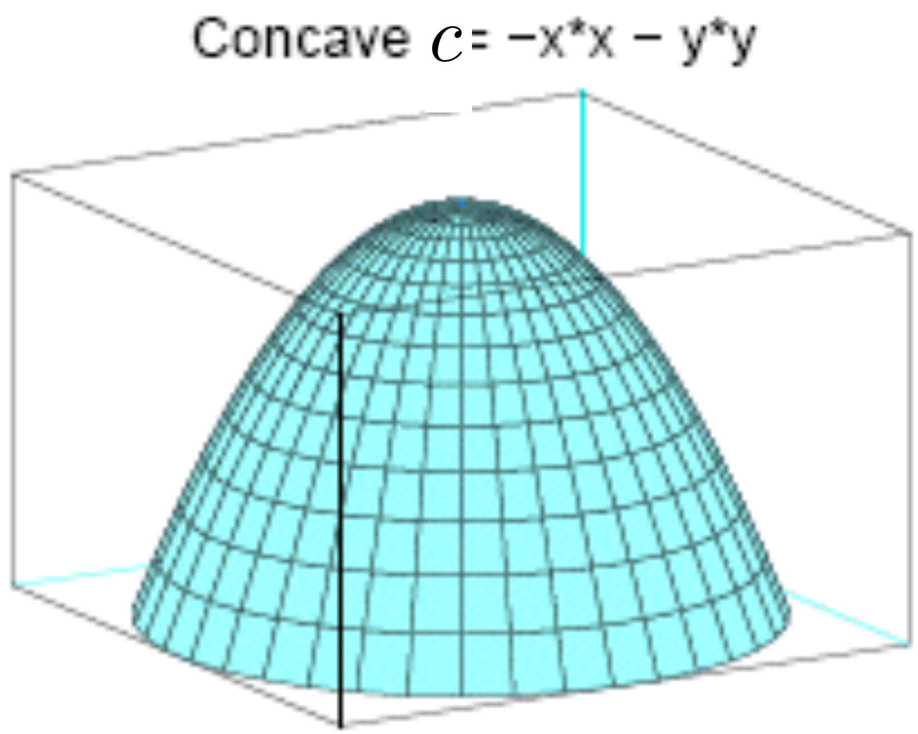
Both have derivative zero under same condition

regardless of $u > 0$

Can either minimize or maximize

$$\arg\min_{\theta} c(\theta) = \arg\max_{\theta} -c(\theta)$$





$$c(t\mathbf{w}_1 + (1-t)\mathbf{w}_2) \leq tc(\mathbf{w}_1) + (1-t)c(\mathbf{w}_2)$$

Reminders and Questions

- Cov[X,X] = V[X], and we also wrote Cov[X] in the assignment
- Datasets for mini-project
  - UCI repository
  - Kaggle competitions
  - Energy datasets (from Prof Omid)
- Today going to go through several examples of ML and MAP
  - These will also be like probability exercises

General format for thought questions

- (1) First show/explain how you understand a concept
- (2) Given this context, propose a follow-up question
- (3) [Optional] Propose an answer to the question, or how you might find it
- Additional note: framing a coherent, concise thought is a skill. When writing your thought question, ask yourself: is this clear?
- Introductory slides have some examples; a few more listed here

Examples of “good” thought questions

- After reading about independence, I wonder how one could check in practice if two variables are independent, given a database of samples? Is this even possible? One possible strategy could be to approximate their conditional distributions, and examine the effects of changing a variable. But it seems like there could be other more direct or efficient strategies.

Examples of “good” thought questions

- “After reading about the definition of expectation and variance, I wonder about a practical real-life problem—when facing with a precious data set with numerous of missing values, how to deal those missing values without causing a huge deviation of expectation and variance of the original data set? One possible scheme is to replace these missing values with expectation (or mean), but it raises up another problem—replacing all the missing values with mean may reduce the variance, causing a huge deviation on the variance of the original dataset.”

Examples of “bad” thought questions

- I don’t understand random variables. Could you explain it again? (i.e. a request for me to explain something, without any insight)

- Derive the maximum likelihood approach for a Gaussian. (i.e., an exercise question from a textbook)

- What is the difference between a probability mass function and a probability density function? (i.e., a question that could easily be answered from reading the definitions in the notes)
    - But the following modification would be good: “I can see that pmfs and pdfs are different, in that the first is for discrete RVs and the second for continuous. Is there a large difference in which one we choose to model our data? Is it sometimes beneficial to discretize continuous variables and use pmfs instead?”

Example of thought questions

- **“Bad” thought question:** I still do not understand what a model is. Are they distributions? Can they be other things?

- **Alternative:** The notion of a model appears to be somewhat imprecise. We have used distributions as a model of our data, with parameters to those distributions representing the model. But, can other thing be models? For example, is plotting the data points and understanding its behavior considered a “model” of the data? What other kinds of models are there?
  - First showed that understood how we have been describing models
  - Then showed follow-up thought about what the term “model” could really mean

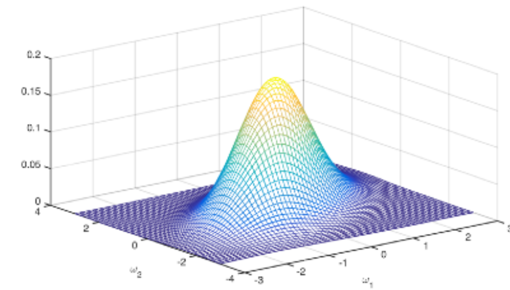
Why this focus on thought questions?

- Whether academia or industry, specifying projects involves understanding what exists, and proposing the “next” thing

- This includes identifying
  - current assumptions/beliefs that could be challenged
  - gaps in current approaches (practical/theoretical)
  - limitations, so can keep those limitations in mind for the solution
  - novel ways forward, given the current solutions/understanding

Clarifications on multivariate Gaussian

$$p(\boldsymbol{\omega}) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{\omega}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\omega}-\boldsymbol{\mu})\right)$$



$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad \boldsymbol{\Sigma} = \begin{bmatrix} 10 & 0 \\ 0 & 2 \end{bmatrix} \quad \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \frac{1}{10} & 0 \\ 0 & \frac{1}{2} \end{bmatrix}$$

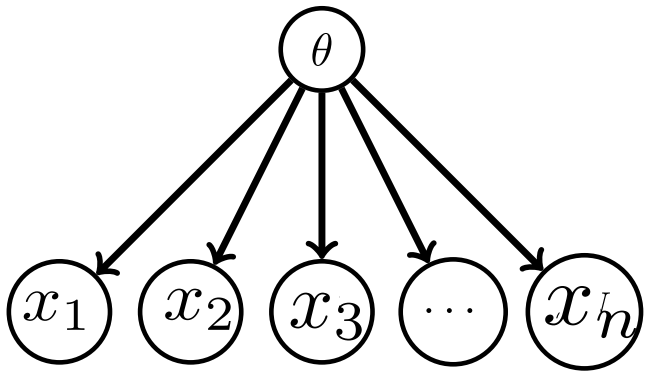
$$\boldsymbol{\omega} - \boldsymbol{\mu} = \begin{bmatrix} \omega_1 - \mu_1 \\ \omega_2 - \mu_2 \end{bmatrix}$$

$$\begin{bmatrix} \omega_1 - \mu_1 \\ \omega_2 - \mu_2 \end{bmatrix} \begin{bmatrix} \frac{1}{10} & 0 \\ 0 & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{10}(\omega_1 - \mu_1) \\ \frac{1}{2}(\omega_2 - \mu_2) \end{bmatrix}$$

$$\begin{bmatrix} \frac{1}{10}(\omega_1 - \mu_1) \\ \frac{1}{2}(\omega_2 - \mu_2) \end{bmatrix}^\top \begin{bmatrix} \omega_1 - \mu_1 \\ \omega_2 - \mu_2 \end{bmatrix} = \frac{1}{10}(\omega_1 - \mu_1)^2 + \frac{1}{2}(\omega_2 - \mu_2)^2$$

Example: maximum likelihood for discrete distributions

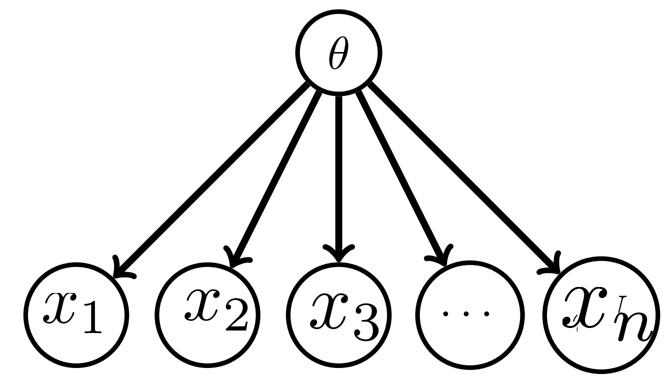
- Imagine you are flipping a biased coin; the model parameter is the bias of the coin, theta
- You get a dataset D = {x_1, …, x_n} of coin flips, where x_i = 1 if it was heads, and x_i = 0 if it was tails
- What is p(D | theta)?



$$
\begin{aligned}
p(\mathcal{D}|\theta) &= p(x_1, \ldots, x_n|\theta) \\
&= \prod_{i=1}^{n} p(x_i|\theta)
\end{aligned}
$$

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

Example: maximum likelihood for discrete distributions



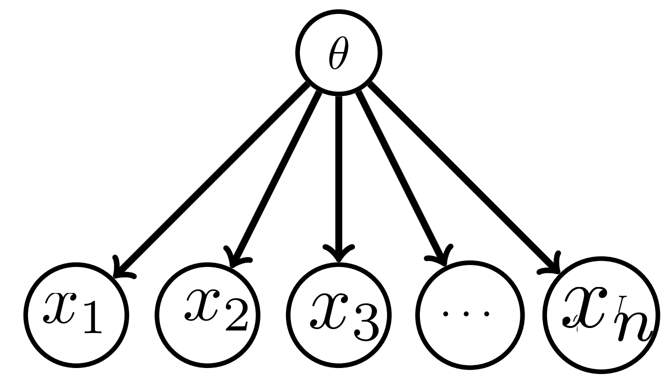
- How do we estimate theta?
- Counting:
  - count the number of heads Nh
  - count the number of tails Nt
  - normalize: theta = Nh/ (Nh + Nt)
- What if you actually try to maximize the likelihood?
  - i.e., solve argmax p(D l theta)

$$
\begin{aligned}
p(\mathcal{D}|\theta) &= p(x_1, \ldots, x_n|\theta) \\
&= \prod_{i=1}^{n} p(x_i|\theta)
\end{aligned}
$$

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

Example: maximum likelihood for discrete distributions

- What if you actually try to maximize the likelihood to get theta?
  - i.e., solve argmax p(D I theta)



$$\max_\theta \prod_{i=1}^{n} p(x_i|\theta) = \max_\theta c(\theta)$$

$$c(\theta) = \prod_{i=1}^{n} p(x_i|\theta)$$

$$\arg\max_\theta c(\theta) = \arg\max_\theta \log c(\theta)$$

$$\begin{aligned} p(\mathcal{D}|\theta) &= p(x_1, \ldots, x_n|\theta) \\ &= \prod_{i=1}^{n} p(x_i|\theta) \end{aligned}$$

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

Example: maximum likelihood for discrete distributions

$$\arg\max_\theta c(\theta) = \arg\max_\theta \log c(\theta)$$

$$\log(ab) = \log a + \log b$$

$$\log(a^c) = c \log a$$

$$\log c(\theta) = \log \prod_{i=1}^{n} p(x_i|\theta)$$

$$= \sum_{i=1}^{n} \log p(x_i|\theta)$$

$$p(x|\theta) = \theta^x (1-\theta)^{1-x}$$

$$\log p(x|\theta) = \log\left(\theta^x\right) + \log\left((1-\theta)^{1-x}\right)$$

$$= x \log(\theta) + (1-x)\log(1-\theta)$$

Example: maximum likelihood for discrete distributions

$$\sum_{i=1}^{n} \log p(x_i|\theta) = \sum_{i=1}^{n} x_i \log(\theta) + \sum_{i=1}^{n} (1 - x_i) \log(1 - \theta)$$

$$= \log(\theta) \left(\sum_{i=1}^{n} x_i \right) + \log(1 - \theta) \left(\sum_{i=1}^{n} (1 - x_i) \right)$$

$$\bar{x} = \sum_{i=1}^{n} x_i$$

$$\frac{d}{d\theta} = \frac{1}{\theta}\bar{x} - \frac{1}{1 - \theta}(n - \bar{x}) = 0$$

Example: maximum likelihood for discrete distributions

$$\bar{x} = \sum_{i=1}^{n} x_i$$

$$\frac{d}{d\theta} = \frac{1}{\theta}\bar{x} - \frac{1}{1-\theta}(n - \bar{x}) = 0$$

$$\implies \frac{\bar{x}}{\theta} = \frac{n - \bar{x}}{1 - \theta}$$

$$\implies (1-\theta)\bar{x} = \theta(n - \bar{x})$$

$$\implies \bar{x} - \theta\bar{x} = \theta n - \theta\bar{x}$$

$$\implies \theta = \frac{\bar{x}}{n}$$

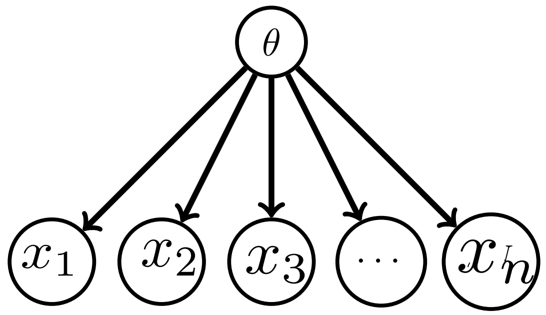
Back to Independence and Conditional independence

- What does it mean to say X and Y are independent?
    - To say “independent”, you have to have a distribution
    - $p_{XY}(x,y) = p_X(x)\, p_Y(y)$, need three three functions
- What is $p_{X|Z}(x | z)$?
    - It's a Bernoulli
- If we know X is a coin, but don’t know the bias, what is p_X?
    - $p_X = \sum_z p_{X|Z}(x | z)\, p(z)$
- What is $p_{XY}(x,y)$?
    - $p_{XY} = \sum_z p_{X,Y|Z}(x,y | z)\, p(z) = \sum_z p_{X|Z}(x | z)\, p_{Y|Z}(y | z)\, p(z)$ not necessarily equal to $p_X(x)\, p_Y(y)$

Example: MAP for discrete distributions

- Imagine you are flipping a biased coin; the model parameter is the bias of the coin, theta

- You get a dataset D = {x_1, …, x_n} of coin
  1 if it was heads, and x_i = 0 if it was tails



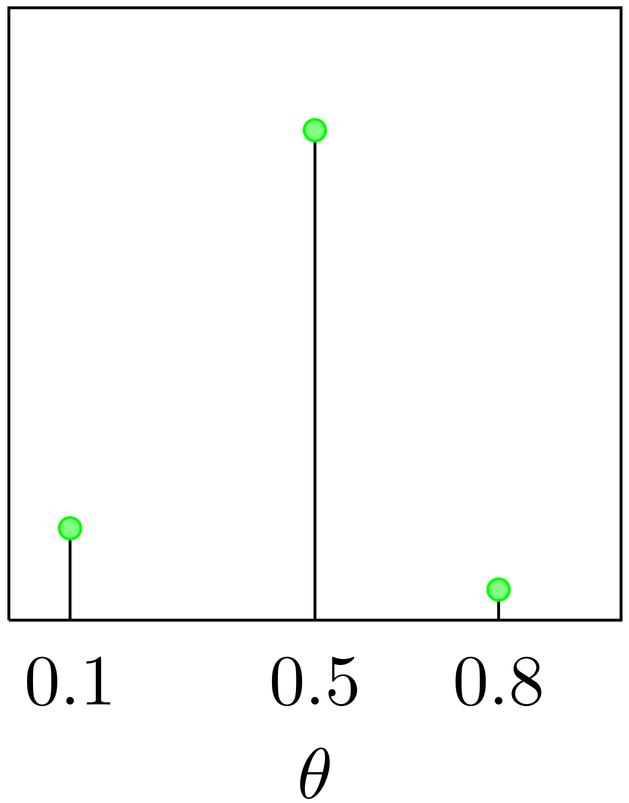
- What if we also specify p(theta)?

- What is the MAP estimate?

Example: MAP for discrete distributions

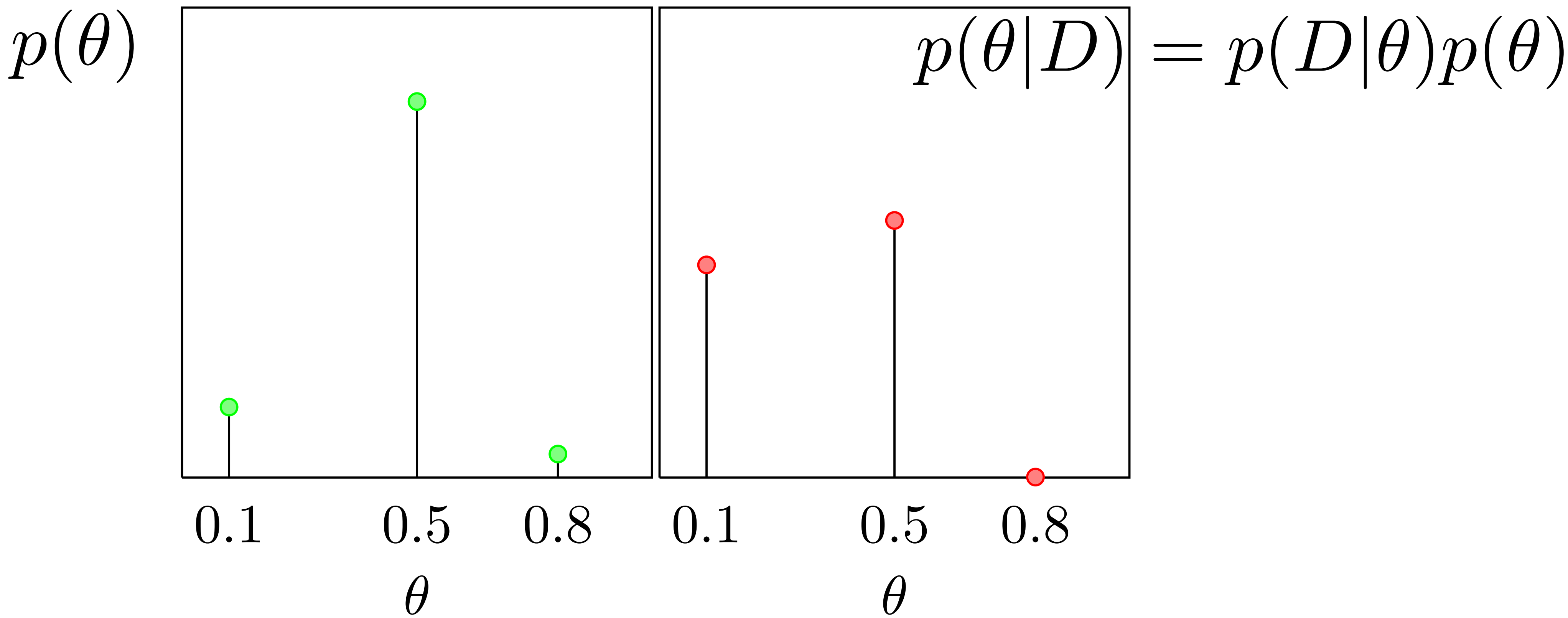
We still need to fully specify the prior $p(\theta)$. To avoid complexities resulting from continuous variables, we'll consider a discrete θ with only three possible states, $\theta \in \{0.1, 0.5, 0.8\}$. Specifically, we assume

$$p(\theta = 0.1) = 0.15, \; p(\theta = 0.5) = 0.8, \; p(\theta = 0.8) = 0.05$$



Example: MAP for discrete distributions

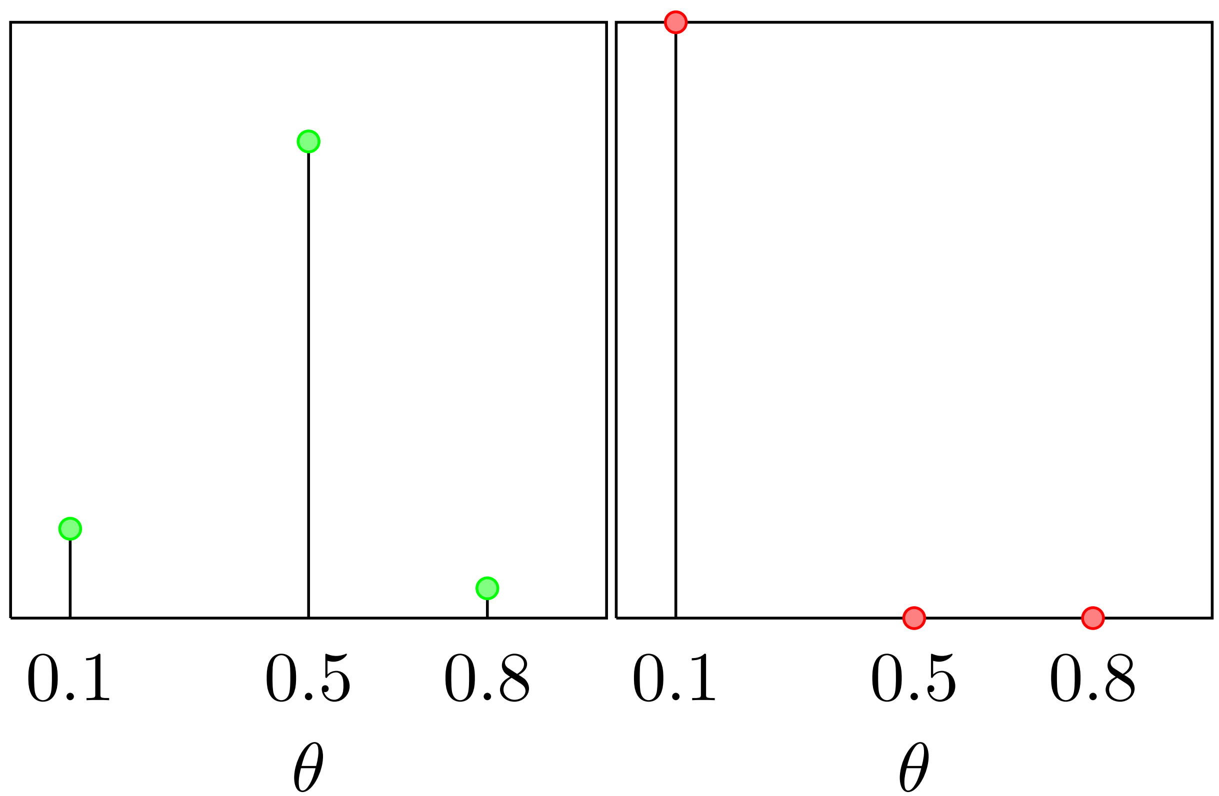
For an experiment with $N_H = 2$, $N_T = 8$, the posterior distribution is



If we were asked to choose a single *a posteriori* most likely value for θ, it would be $\theta = 0.5$, although our confidence in this is low since the posterior belief that $\theta = 0.1$ is also appreciable. This result is intuitive since, even though we observed more Tails than Heads, our prior belief was that it was more likely the coin is fair.

Example: MAP for discrete distributions

Repeating the above with $N_H = 20$, $N_T = 80$, the posterior changes to



so that the posterior belief in $\theta = 0.1$ dominates. There are so many more tails than heads that this is unlikely to occur from a fair coin. Even though we *a priori* thought that the coin was fair, *a posteriori* we have enough evidence to change our minds.

Now on to some careful examples of MAP!

- Whiteboard for Examples 8, 9, 10, 11

- More fun with derivatives and finding the minimum of a function

- Next class:
    - finish off parameter estimation
    - introduction to prediction problems for ML